Management (Montevideo). 2025; 3:206
doi: 10.62486/agma2025206
ISSN: 3046-4048

ORIGINAL

E-commerce web system with personalized recommendations based on purchase history for American Eagle in Maná, Ecuador

Sistema web de Comercio Electrónico con recomendaciones personalizadas basadas en el historial de compras para la empresa American Eagle en la Maná, Ecuador

Jennifer Valeria Faz Gallardo[1], Valeria Anabel Guaman Capa[1], Rodolfo Najarro Quintero[1]

[1]Universidad Técnica de Cotopaxi UTC, Extensión La Maná, Latacunga, Ecuador.

Citar como: Faz Gallardo JV, Guaman Capa VA, Najarro Quintero R. E-commerce web system with personalized recommendations based on purchase history for American Eagle in Maná, Ecuador. Management (Montevideo). 2025; 3:206. https://doi.org/10.62486/agma2025206

Enviado: 28-03-2025 **Revisado:** 12-07-2025 **Aceptado:** 14-10-2025 **Publicado:** 15-10-2025

Editor: Ing. Misael Ron

Autor para la correspondencia: Jennifer Valeria Faz Gallardo

ABSTRACT

Introduction: in Ecuador, fashion retailers face volatile catalogs and low-density transaction records, which hinders the personalization of the customer experience. This study aimed to evaluate whether a recommendation engine based solely on purchase history improves performance and user experience for the American Eagle store in La Maná.
Method: a web-based e-commerce system was implemented using an MVC architecture and a two-stage framework: candidate generation using co-occurrences and collaborative item-to-item filtering (cosine/Jaccard), followed by ordering with BPR-MF based on implicit feedback. The offline evaluation included ranking metrics (Precision@k, Recall@k, MAP, NDCG@k), coverage, and diversity. Uncertainty was calculated with 95 % confidence intervals using bootstrap, and significance was calculated using the paired Wilcoxon test. Popularity, Item-CF, and User-CF were used as references. In addition, an A/B test was applied measuring CTR, search time and usability (SUS/UEQ).
Results: the proposed system showed consistent improvements in NDCG@10 and Recall@10 compared to the baseline models, maintained comparable Precision@10, and increased both catalog coverage and diversity. In user testing, it increased CTR, reduced search time, and achieved favorable usability.
Conclusions: a purely historical recommendation approach is viable for SMEs with moderate resources, improves product discovery, and supports evidence-based merchandising decisions. The study provides a replicable protocol applicable to other contexts and seasons in Latin America.
or unstructured, with a length of no more than 250 words; written in the past tense and in the third person singular.

Keywords: E-Commerce; Recommendation Systems; Purchase History; BPR-MF; NDCG; Coverage; Diversity.

RESUMEN

Introducción: en Ecuador, los comercios de moda enfrentan catálogos volátiles y registros transaccionales de baja densidad, lo que dificulta la personalización de la experiencia del cliente. Este estudio tuvo como objetivo evaluar si un motor de recomendaciones basado únicamente en el historial de compras mejora el desempeño y la experiencia de usuario en el caso de la tienda American Eagle en La Maná.
Método: se implementó un sistema web de comercio electrónico con arquitectura MVC y un esquema de dos etapas: generación de candidatos mediante co-ocurrencias y filtrado colaborativo ítem-ítem (coseno/Jaccard), seguido de ordenamiento con BPR-MF basado en feedback implícito. La evaluación offline incluyó

métricas de ranking (Precision@k, Recall@k, MAP, NDCG@k), cobertura y diversidad. La incertidumbre se calculó con intervalos de confianza al 95 % mediante bootstrap, y la significancia con prueba de Wilcoxon pareado. Como referencias se usaron popularidad, Item-CF y User-CF. Además, se aplicó una prueba A/B midiendo CTR, tiempo de búsqueda y usabilidad (SUS/UEQ).
Resultados: el sistema propuesto mostró mejoras consistentes en NDCG@10 y Recall@10 respecto a los modelos base, mantuvo Precision@10 comparable y elevó tanto la cobertura como la diversidad del catálogo. En la prueba con usuarios, aumentó el CTR, redujo el tiempo de búsqueda y alcanzó una usabilidad favorable.
Conclusiones: un enfoque de recomendaciones puramente histórico resulta viable en pymes con recursos moderados, mejora el descubrimiento de productos y apoya decisiones de merchandising basadas en evidencia. El estudio aporta un protocolo reproducible aplicable a otros contextos y temporadas en Latinoamérica.
o no estructurado, con una extensión no mayor a 250 palabras; redactado en pasado y en tercera persona del singular.

Palabras clave: Comercio Electrónico; Sistemas de Recomendación; Historial de Compras; BPR-MF; NDCG; Cobertura; Diversidad.

INTRODUCCIÓN

La personalización mediante sistemas de recomendación basados en el historial de compras constituye un eje fundamental del comercio electrónico contemporáneo, particularmente en el sector del retail de moda, caracterizado por su alta rotación de catálogos, marcada estacionalidad y la frecuente dispersión de interacciones. En este ámbito, los motores de recomendación históricos se han consolidado como soluciones de referencia, implementados bajo arquitecturas de dos etapas centradas en datos transaccionales. Dicho enfoque consiste en la generación de candidatos a través de co-ocurrencias y técnicas de filtrado colaborativo, tanto basadas en usuarios como en ítems, seguida de un ordenamiento mediante métodos de ranking apoyados en retroalimentación implícita, entre los cuales se destacan algoritmos como BPR-MF.[1,2,3]

La literatura especializada ha demostrado que esta aproximación mejora la relevancia de las recomendaciones y favorece el descubrimiento de productos en escenarios de catálogos altamente dinámicos.[4,5] Sin embargo, pese a sus avances, el campo enfrenta aún desafíos metodológicos de gran relevancia, entre ellos la necesidad de lograr un equilibrio entre precisión, cobertura del catálogo y diversidad de resultados, además de asegurar la adecuada calibración de las recomendaciones al perfil particular de cada usuario. En consecuencia, resulta indispensable complementar las métricas clásicas de ranking con indicadores más allá de la exactitud, tales como la diversidad intra-lista o la cobertura del catálogo, incorporando además evaluaciones estadísticas rigurosas mediante intervalos de confianza y pruebas no paramétricas que otorguen robustez a los hallazgos y sustenten decisiones de despliegue en entornos reales de producción.[4,6,7,8]

En el sector de la moda, estudios recientes han señalado particularidades que complejizan el diseño de sistemas de recomendación, como la compatibilidad de atuendos, la necesidad de integrar atributos visuales y textuales, y el problema de cold-start derivado de tendencias efímeras.[7,8,9,10] Estas revisiones subrayan la pertinencia de enfoques basados en historial debidamente instrumentados, co-ocurrencias, filtrado colaborativo por vecindad y factorización con feedback implícito, que, al complementarse con criterios de catálogo, permiten sostener la cobertura y la diversidad sin comprometer la precisión. No obstante, se identifican brechas relevantes que limitan la aplicabilidad de estas técnicas en pequeñas y medianas empresas (pymes) de mercados emergentes. Entre ellas, destaca la escasez de estudios empíricos que documenten de forma detallada las restricciones operativas reales, tales como la densidad de datos, la rotación de inventario o los sesgos de popularidad, y que al mismo tiempo ofrezcan protocolos reproducibles. Asimismo, muchas evaluaciones se han limitado al análisis de métricas de precisión, dejando de lado indicadores de diversidad y cobertura, o la incorporación de análisis de significancia estadística que permitan extraer conclusiones sólidas a nivel de usuario. A esto se suma la calibración insuficiente de las recomendaciones al perfil de intereses del consumidor, problemática que persiste incluso frente a propuestas canónicas y desarrollos recientes orientados a optimizar listas Top-N.[7]

Frente a estas brechas, el presente estudio propone la evaluación de un sistema de recomendación puramente histórico en un caso real correspondiente a la tienda American Eagle en La Maná, Ecuador. El objetivo central es valorar rigurosamente el desempeño del motor bajo un protocolo de evaluación reproducible que integre métricas de ranking, como NDCG y Recall, indicadores de catálogo relacionados con la cobertura y la diversidad, y validaciones directas con usuarios finales a través de experimentos controlados A/B. Las preguntas de investigación se enfocan en determinar en qué medida el sistema supera a los baselines tradicionales de popularidad, filtrado colaborativo por ítems y por usuarios; cuál es su impacto en la cobertura y diversidad del catálogo; si las mejoras observadas en pruebas offline se traducen efectivamente en beneficios tangibles para los consumidores, tales como un mayor CTR, menor tiempo de búsqueda y mejor percepción de usabilidad; y

cómo influyen factores como el tamaño del Top-N o la longitud del historial en el equilibrio entre precisión y diversidad.

Las contribuciones de este trabajo son múltiples y de especial valor para el contexto latinoamericano.

En primer lugar, se presenta el diseño e implementación de un pipeline de dos etapas basado exclusivamente en historial de compras, documentando de manera transparente los supuestos, divisiones temporales e hiperparámetros para garantizar la reproducibilidad. En segundo lugar, se ofrece un protocolo de evaluación que integra métricas de ranking y de catálogo con análisis estadísticos rigurosos, alineado con los marcos más recientes de validación en sistemas de recomendación. En tercer lugar, se aporta evidencia empírica obtenida en un entorno real mediante pruebas A/B y mediciones de usabilidad, lo cual fortalece la validez de constructo de los resultados. Adicionalmente, se realizan experimentos de ablación y análisis de sensibilidad sobre los distintos componentes del sistema, discutiendo los compromisos entre precisión y diversidad y la calibración de la mezcla de categorías recomendadas. Finalmente, se destaca un enfoque de transparencia y gobernanza de datos, que incluye la anonimización, la política de retención, la autorización para el uso del nombre comercial y la publicación de recursos reproducibles con DOI, a fin de facilitar la transferencia de este modelo a otras pymes de mercados emergentes.

MÉTODO

La metodología del presente estudio se estructuró para garantizar rigor científico, reproducibilidad y aplicabilidad práctica en el contexto de un retailer de moda como American Eagle en La Maná, cuya operación combina presencia física y canal de comercio electrónico. La base de datos transaccional utilizada registra, para cada interacción, el identificador de usuario, el identificador del ítem, la marca temporal, la cantidad, el precio y los metadatos asociados al producto, tales como categoría, talla y color. A partir de estas secuencias, se construyeron historiales de compra por usuario, siguiendo un proceso de depuración exhaustiva que incluyó la eliminación de duplicados, la detección de bots y sesiones anómalas con base en reglas de velocidad y patrones de clic o compra, así como la unificación de la taxonomía de categorías y la normalización de valores atípicos en cantidad y precio. Todos los identificadores personales fueron anonimizados mediante algoritmos de hash irreversible y claves sustitutas, y los datos se gestionaron bajo políticas de acceso controlado y retención previamente definidas, cumpliendo con criterios éticos y normativos.

Para mejorar la calidad de los datos se establecieron criterios de inclusión, considerando únicamente usuarios con un número mínimo de interacciones válidas y artículos con umbral de ventas o visualizaciones, de modo que la muestra analizada representara un comportamiento significativo y no sesgado por interacciones marginales. Posteriormente, se aplicó una partición temporal estricta que dividió el conjunto de datos en entrenamiento, validación y prueba, con cortes cronológicos que evitaron la fuga de información. En el conjunto de prueba se utilizó un esquema leave-one-out por usuario, tomando su última interacción como objetivo a predecir, lo que garantizó un diseño realista y alineado con el comportamiento secuencial del consumo.

El sistema propuesto fue implementado siguiendo un patrón two-stage puramente histórico, diseñado para priorizar baja latencia, reproducibilidad y trazabilidad de versiones.

En la primera etapa, denominada generación de candidatos, se construyeron listas preliminares combinando dos enfoques complementarios. Por un lado, se calcularon co-ocurrencias a nivel de sesión u orden, ponderadas con un decaimiento temporal que daba mayor relevancia a interacciones recientes; por otro, se empleó filtrado colaborativo ítem-ítem sobre la matriz usuario–ítem, utilizando medidas de similitud como coseno y Jaccard, lo que permitió recuperar los vecinos más cercanos de cada ítem y enriquecer el conjunto de candidatos. Además, para afrontar el problema de arranque en frío y los cambios derivados de la estacionalidad, se incorporó una heurística de popularidad condicionada por segmento de producto y temporada. En todos los casos, se excluyeron ítems no disponibles o ya adquiridos por el usuario, aplicándose también filtros de negocio relacionados con talla, color y stock.

La segunda etapa correspondió al ordenamiento de candidatos mediante Bayesian Personalized Ranking con Factorización Matricial (BPR-MF), optimizando sobre pares de interacciones positivas y negativas construidos a partir de feedback implícito. Este modelo se entrenó utilizando muestreo negativo, regularización L2 y el optimizador Adam, estableciendo criterios de early stopping según el desempeño en validación. La dimensión latente de los factores se ajustó de acuerdo con experimentos de sensibilidad y el score final de ranking se definió como el producto interno de los vectores latentes de usuario e ítem. Para reforzar la trazabilidad metodológica, se fijaron semillas, se documentaron las particiones temporales y se registraron todas las configuraciones e hiperparámetros empleados.

Como línea base de comparación se consideraron tres enfoques tradicionales: popularidad global y segmentada, filtrado colaborativo basado en usuarios y filtrado colaborativo basado en ítems. Estos permitieron contrastar el desempeño del sistema propuesto frente a modelos simples, pero ampliamente utilizados en la industria. Adicionalmente, se diseñaron variantes y pruebas de ablación que evaluaron el impacto de los diferentes componentes del pipeline, tales como co-ocurrencias, Item-CF, la inclusión del módulo de ranking

BPR-MF, la longitud del historial por usuario y el tamaño del Top-N recomendado. También se introdujeron reglas de negocio específicas, como la obligatoriedad de considerar stock y la imposición de una diversidad mínima en las listas, con el fin de analizar cómo tales políticas influían en el equilibrio entre precisión y cobertura.

La evaluación se realizó mediante métricas de ranking, como Precisión, Recall, MAP y NDCG a diferentes valores de k, y métricas de catálogo, incluyendo cobertura, diversidad intra-lista y novedad. Todas estas métricas fueron calculadas por usuario y posteriormente agregadas a nivel macro, empleando tanto la media como, cuando fue pertinente, la mediana. La estimación de incertidumbre se efectuó a través de bootstrap estratificado con múltiples réplicas, obteniéndose intervalos de confianza al 95 %. Las comparaciones entre el sistema propuesto y los modelos de referencia se realizaron mediante pruebas no paramétricas de Wilcoxon para datos apareados, con ajuste por multiplicidad en los contrastes múltiples.

Como complemento, se ejecutó un experimento A/B en el canal de comercio electrónico, con asignación aleatoria de usuarios a grupo control y experimental, estratificado según el tipo de cliente y categoría de producto. En este entorno se midieron métricas primarias como el CTR y el tiempo de búsqueda, así como métricas secundarias relacionadas con conversión y tasa de rebote, incorporándose también métricas de estabilidad operacional como latencia y errores de servidor.

La percepción de usabilidad fue evaluada mediante los cuestionarios estandarizados SUS y UEQ, y los resultados se analizaron con estadística descriptiva, pruebas de significancia y estimación de tamaños de efecto. Finalmente, se establecieron procedimientos claros de reproducibilidad que incluyeron repositorios con scripts, configuraciones, documentación detallada y datos sintéticos anonimizados, lo cual asegura la transferibilidad del protocolo a otras pymes en contextos similares. La validez del estudio se reforzó con estrategias para mitigar amenazas internas, externas, de constructo y de conclusión, garantizando un marco metodológico sólido y confiable.

RESULTADOS

Desempeño offline (ranking y catálogo)

El análisis offline se centró en las métricas de ranking (Precision@k, Recall@k, MAP@k, NDCG@k) y de catálogo (cobertura, diversidad intra-lista y novelty), evaluadas en distintos tamaños de recomendación (k = 5, 10, 20). Las Tablas 1A y 1B presentan los resultados promedio por usuario, expresados como media ± IC95 %, obtenidos mediante bootstrap estratificado. Para las comparaciones pareadas se empleó la prueba de Wilcoxon (α = 0,05), confirmando la existencia de diferencias estadísticamente significativas entre el sistema basado en historial y los modelos de referencia.

La tabla 1 muestra los valores de NDCG y Recall, donde el sistema histórico (co-ocurrencias + BPR-MF) alcanzó mejoras consistentes frente a los baselines. En particular, para NDCG@10 y Recall@10 se obtuvieron incrementos notables respecto a Item-CF y User-CF, mientras que la ganancia frente al modelo de popularidad fue aún más pronunciada. Estos resultados indican que la combinación de co-ocurrencias con BPR-MF logra capturar patrones más representativos del comportamiento del usuario, generando recomendaciones de mayor relevancia.

Tabla 1. Desempeño offline — NDCG@k y Recall@k (media por usuarios ± IC95 %)

Modelo	NDCG@5 (±IC95 %)	Recall@5 (±IC95 %)	NDCG@10 (±IC95 %)	Recall@10 (±IC95 %)	NDCG@20 (±IC95 %)	Recall@20 (±IC95 %)
Popularidad	0,142 (0,137-0,147)	0,062 (0,058-0,066)	0,136 (0,132-0,140)	0,110 (0,105-0,115)	0,128 (0,125-0,131)	0,184 (0,178-0,190)
Item-CF	0,188 (0,182-0,194)	0,077 (0,073-0,081)	0,180 (0,175-0,185)	0,165 (0,160-0,170)	0,168 (0,164-0,172)	0,283 (0,276-0,290)
User-CF	0,178 (0,172-0,184)	0,073 (0,069-0,077)	0,171 (0,166-0,176)	0,156 (0,151-0,161)	0,160 (0,156-0,164)	0,270 (0,263-0,277)
Sistema basado en historial (Co-ocurrencias + BPR-MF)	0,204 (0,198-0,210)	0,086 (0,082-0,090)	0,196 (0,191-0,201)	0,182 (0,176-0,188)	0,184 (0,180-0,188)	0,306 (0,298-0,314)

Fuente. Elaboración propia con datos transaccionales anonimizados de American Eagle en La Maná, período [2021- 03 a 2025-07].
Nota. Métricas de ranking por usuario: NDCG@k y Recall@k para k∈ {5,10,20}. Valores como media ± IC95 % (bootstrap estratificado, B=1000). Comparaciones pareadas con Wilcoxon (α=0,05). Partición temporal (train/valid/test) y evaluación leave-one-out en prueba. Zona horaria: America/Guayaquil.

En la tabla 2 se presentan los indicadores de precisión y métricas de catálogo. Se observa que el sistema propuesto, aunque mantiene niveles de Precisión@k comparables a los baselines, supera de manera significativa en MAP, cobertura y diversidad intra-lista. Destaca el aumento de la cobertura, que pasó de valores de 22,6

% en el modelo de popularidad a 41,5 % en el sistema basado en historial, lo que implica una exploración más amplia del catálogo disponible. De forma similar, la diversidad intra-lista (ILD@10) y la novedad presentaron incrementos relevantes, evidenciando que las recomendaciones no solo fueron más precisas, también más variadas y con menor sesgo hacia ítems de alta popularidad.

Tabla 2. Desempeño offline — Precision@k, MAP@k y métricas de catálogo (media por usuario ± IC95 %)

Modelo	Precision@5 (±IC95 %)	MAP@5 (±IC95 %)	Precision@10 (±IC95%)	MAP@10 (±IC95 %)	Precision@20 (±IC95%)	MAP@20 (±IC95 %)	Coverage (%)	ILD@10	aNovelty@10
Popularidad	0,118 (0,112-0,124)	0,091 (0,087-0,095)	0,102 (0,098-0,106)	0,086 (0,082-0,090)	0,086 (0,083-0,089)	0,078 (0,075-0,081)	22,6 % (20,8-23,2)	0,520 (0,510-0,530)	2,880 (2,820-2,940)
Item-CF	0,145 (0,140-0,150)	0,114 (0,110-0,118)	0,124 (0,120-0,128)	0,106 (0,102-0,110)	0,104 (0,101-0,107)	0,096 (0,093-0,099)	33,4 % (32,0-34,8)	0,660 (0,650-0,670)	3,830 (3,770-3,890)
User-CF	0,137 (0,132-0,142)	0,108 (0,104-0,112)	0,118 (0,114-0,122)	0,101 (0,097-0,105)	0,100 (0,097-0,103)	0,092 (0,089-0,095)	29,7 % (28,4-31,0)	0,640 (0,630-0,650)	3,760 (3,700-3,820)
Sistema basado en historial (Co-ocurrencias +BPR-MF)	0,162 (0,156-0,168)	0,127 (0,122-0,132)	0,139 (0,135-0,143)	0,118 (0,114-0,122)	0,118 (0,115-0,121)	0,107 (0,104-0,110)	41,5 % (40,0-43,0)	0,720 (0,710-0,730)	4,100 (4,030-4,170)

Fuente. Elaboración propia con datos transaccionales anonimizados de American Eagle en La Maná, período [2021- 03 a 2025-07].
Nota. Ranking: Precision@k y MAP@k para k∈{5,10,20}. Catálogo: Cobertura = |∪u T^u_k|/|I_disponible|; ILD@10 = disimilitud promedio intra-lista; Novelty@10 = −(1/10) $\sum \log_2 p(i)$. Valores como media ± IC95 % (bootstrap estratificado, B=1000). Wilcoxon pareado (α=0,05) con ajuste por multiplicidad cuando aplica.

Los resultados offline sugieren que un enfoque puramente histórico puede equilibrar adecuadamente la relevancia con la diversidad, mitigando las limitaciones de los sistemas tradicionales que tienden a favorecer únicamente ítems populares.

Prueba con usuarios (A/B)

La validación online se llevó a cabo mediante un experimento A/B en el canal de comercio electrónico de American Eagle en La Maná. La muestra estuvo compuesta por más de 10 000 sesiones, distribuidas en dos grupos balanceados (control = 5 120 sesiones; tratamiento = 5 148 sesiones), con estratificación por tipo de cliente (nuevo vs. recurrente). El ensayo tuvo una duración de 14 días, permitiendo cubrir distintos patrones de tráfico y demanda.

La tabla 3 sintetiza los resultados. En términos de CTR, el sistema basado en historial incrementó el indicador en 2,2 puntos porcentuales frente al control (p = 0,004), lo que refleja un mayor atractivo y pertinencia de las recomendaciones. El tiempo promedio de búsqueda se redujo en 7,2 segundos (p = 0,006), mostrando que los usuarios localizaron más rápidamente productos de interés. Además, se observó un aumento en la conversión micro ("añadir al carrito"), con una diferencia absoluta de +1,2 p.p. (p = 0,018). En cuanto a la percepción subjetiva de usabilidad, las mejoras fueron consistentes: el puntaje SUS pasó de 76,2 a 82,8 (Δ = +6,6; p = 0,011) con alta confiabilidad interna (α = 0,86), mientras que el índice global UEQ también presentó un incremento significativo (+0,30; p = 0,021).

Tabla 3. Ensayo A/B y usabilidad — resultados (media por usuario/sesión ± IC95 %)

Métrica	Control (IC95 %)	Tratamiento (IC95 %)	Δ Absoluto (IC95 %)	p-valor	α Cronbach
CTR (%)	7,4 % (6,9–7,9)	9,6 % (9,1–10,2)	+2,2 p.p. (+1,1–+3,3)	0,004	
Tiempo de búsqueda (s)	42,8 (40,1–45,5) s	35,6 (33,0–38,2) s	−7,2 (−10,0 a −4,3) s	0,006	
Añadir al carrito (%)	4,9 % (4,5–5,3)	6,1 % (5,6–6,6)	+1,2 p.p. (+0,4–+2,0)	0,018	
SUS (0–100)	76,2 (74,0–78,4)	82,8 (81,0–84,6)	+6,6 (+3,8–+9,4)	0,011	0,86
UEQ (global, −3 a +3)	1,1 (1,0–1,3)	1,4 (1,3–1,6)	+0,30 (+0,12–+0,48)	0,021	0,88

Fuente. Elaboración propia con datos experimentales anonimizados de American Eagle en La Maná, período [2021- 03 a 2025-07].
Nota. CTR y Añadir al carrito: prueba de diferencia de proporciones bilateral, con IC95 % por bootstrap. Tiempo de búsqueda: Mann–Whitney U (o t de Welch si procede). SUS/UEQ: medias con IC95 %; α de Cronbach como confiabilidad interna. IC95 % estimados por bootstrap estratificado (B=1000) a nivel de usuario/sesión. Métricas guardarraíl (latencia de página, errores 4xx/5xx) se mantuvieron sin cambios significativos.

Estos hallazgos confirman que los beneficios observados en el análisis offline se trasladan al comportamiento real de los usuarios, logrando un impacto positivo tanto en métricas objetivas de interacción como en la experiencia percibida.

Robustez y validación cruzada

Las pruebas de robustez demostraron que los efectos del sistema se mantuvieron estables a lo largo de diferentes subventanas temporales, con variaciones relativas dentro de márgenes controlados y diferencias pareadas significativas frente al mejor baseline ($p < 0,05$). Asimismo, el análisis por segmentos de usuario (nuevo vs. recurrente) y por categorías de producto (calzado, tops, accesorios) no evidenció interacciones significativas entre modelo y segmento, lo que sugiere que el sistema conserva su efectividad de manera transversal.

En cuanto a la relación entre métricas offline y desempeño online, se identificó una correlación positiva entre NDCG@10 y CTR, tanto con correlación de Pearson como de Spearman, reforzando la validez de constructo del protocolo de evaluación. Finalmente, la estabilidad del sistema fue confirmada para distintos tamaños de recomendación ($k = 5, 10, 20$), manteniendo un orden inalterado entre modelos y diferencias significativas en los cortes principales.

DISCUSIÓN

Hallazgos principales e interpretación

Los resultados de esta investigación muestran que un enfoque puramente histórico en dos etapas, generación de candidatos mediante co-ocurrencias y filtrado colaborativo ítem-ítem, seguido de un ordenamiento con BPR-MF, supera consistentemente a los modelos de referencia basados en popularidad, Item-CF y User-CF en las métricas de NDCG@k y Recall@k para diferentes valores de k. Las diferencias fueron estadísticamente significativas y se mantuvieron estables en los análisis por segmentos y frente a variaciones en el tamaño de la lista recomendada. Esto sugiere que la señal transaccional contenida en los historiales de compra es suficiente para mejorar de forma sostenida el ordenamiento de recomendaciones en catálogos de moda caracterizados por alta rotación y sparsity.[1,2,3]

Más allá de la exactitud, el sistema evidenció incrementos notables en cobertura y diversidad intra-lista sin un deterioro relevante en Precision@k. Este patrón coincide con la literatura que recomienda integrar indicadores "más allá de la exactitud" para sostener la larga cola de productos y mitigar sesgos de popularidad, manteniendo un balance adecuado entre descubrimiento y relevancia.[6,7,8] La traslación de estos resultados offline al entorno online fue consistente: las mejoras en NDCG@k y Recall@k se reflejaron en mayores tasas de clic (CTR) y menores tiempos de búsqueda en la prueba A/B, reforzando la validez de constructo del protocolo adoptado. La correlación positiva entre NDCG@10 y CTR confirma que los indicadores de ranking son buenos predictores de beneficios tangibles en la interacción con usuarios.[7]

En su conjunto, la evidencia posiciona al pipeline histórico two-stage como una alternativa eficaz y operativamente viable para pymes de moda. Este enfoque aprovecha señales colaborativas de bajo costo computacional, es fácilmente reproducible y se integra de manera natural con prácticas de monitoreo y control estadístico en producción.[1,2,3]

Aportes teóricos y metodológicos

En el plano teórico, este estudio demuestra que un pipeline puramente histórico puede mejorar simultáneamente métricas de ranking (NDCG@k, Recall@k) y de catálogo (cobertura y diversidad) en un contexto real de retail de moda con alta rotación de inventario y datos escasos. Esto establece condiciones de suficiencia para enfoques basados en historial en pymes. Asimismo, operacionaliza un marco de "evaluación más allá de la exactitud", integrando cobertura y diversidad intra-lista como objetivos de primer orden y discutiendo su equilibrio con la precisión del ranking, en línea con la literatura reciente.[6,7,8] Otro aporte teórico relevante es la validación de constructo: se aportan evidencias de correlación positiva entre desempeño offline y métricas de interacción real, reforzando la utilidad de las métricas de ranking como predictoras de beneficios prácticos.[7] Finalmente, se enmarca la calibración del Top-N como criterio explícito en sistemas históricos, apoyándose en formulaciones previas sobre recomendaciones calibradas.[8,9,10]

En el plano metodológico, se presenta un protocolo reproducible que combina particiones temporales estrictas, evaluación leave-one-out, métricas promediadas a nivel de usuario, intervalos de confianza por bootstrap y comparaciones no paramétricas pareadas, lo que fortalece la robustez estadística. Además, se incluyen análisis de sensibilidad y experimentos de ablación centrados en componentes históricos, lo cual permite atribuir efectos y evaluar trade-offs entre precisión y diversidad. También se estandariza la presentación de resultados en bloques claros y complementarios, evitando redundancias y facilitando la auditoría estadística. Por último, se documenta un esquema de gobernanza y trazabilidad que asegura anonimización, control de datos y disponibilidad de recursos reproducibles, alineado con las recomendaciones actuales sobre transparencia en

sistemas de recomendación.[1]

Implicaciones prácticas para pymes de moda

Los hallazgos tienen implicaciones directas para American Eagle en La Maná y para otras pymes del sector. En primer lugar, muestran que una arquitectura two-stage basada en historial puede implementarse con bajo costo computacional, utilizando precálculo nocturno y cachés segmentados, garantizando tiempos de respuesta competitivos. En segundo lugar, sugieren que los indicadores más allá de la exactitud deben incorporarse como parte del tablero operativo de métricas, con umbrales mínimos de cobertura y diversidad para asegurar exposición de la larga cola de productos. En tercer lugar, se enfatiza la importancia de la experimentación continua mediante pruebas A/B con estratificación por segmento de usuario, lo que permite validar el impacto real de las recomendaciones y ajustar parámetros en producción. Finalmente, se subraya la necesidad de fortalecer la gobernanza de datos, incluyendo anonimización, consentimiento y trazabilidad de modelos, así como mantener filtros stock-aware y reglas blandas de merchandising que equilibren los objetivos comerciales con la experiencia de usuario. Los resultados integrados sugieren que, un pipeline histórico bien instrumentado permite a una pyme de moda mejorar simultáneamente precisión, cobertura y diversidad con costos controlados, fortaleciendo tanto la experiencia de usuario como la rotación de inventario.[11,12]

Amenazas a la validez

Como en todo estudio empírico, existen amenazas que deben ser consideradas. En cuanto a la validez interna, los sesgos de popularidad y estacionalidad pudieron influir en el desempeño offline, aunque fueron mitigados con particiones temporales, decaimiento temporal y comparaciones a nivel de usuario. También se controló el riesgo de fuga de información mediante cortes cronológicos y la asignación aleatoria en el A/B, aunque subsisten riesgos de instrumentación. Respecto a la validez de constructo, la elección de métricas pudo favorecer determinados patrones de ranking, aunque se incorporaron indicadores complementarios como cobertura y diversidad para equilibrar la evaluación. La validez externa se limita por tratarse de un caso único en una geografía y dominio específicos, lo que exige replicación en otros contextos. Finalmente, la validez de conclusión se ve condicionada por la multiplicidad de pruebas y el tuning de hiperparámetros; para mitigarlo, se reportaron intervalos de confianza, se aplicaron ajustes por multiplicidad y se vigiló la potencia estadística del experimento A/B.

Líneas futuras

El estudio abre varias líneas de investigación. En primer lugar, explorar enfoques de aprendizaje online y algoritmos tipo bandit que balanceen explotación de historiales con exploración de ítems de la larga cola. En segundo lugar, avanzar en la calibración explícita de listas Top-N mediante objetivos multiobjetivo que integren precisión y diversidad de manera controlada. En tercer lugar, incorporar información multimodal (texto e imágenes) para mejorar la compatibilidad de atuendos y mitigar problemas de cold-start. En cuarto lugar, implementar reranking cost-aware que integre restricciones de latencia y presupuestos computacionales, asegurando estabilidad en escenarios de alta demanda. Finalmente, realizar evaluaciones longitudinales que incluyan métricas de rotación de inventario, fidelización y equidad en la exposición de productos, aportando una visión integral del impacto de los sistemas de recomendación en la sostenibilidad de las pymes de moda.

En síntesis, este estudio confirma que los sistemas históricos, cuando se evalúan bajo marcos rigurosos y se integran con métricas de catálogo y pruebas en usuarios reales, representan una alternativa eficaz, replicable y costo-eficiente para la personalización en pymes de moda. Su consolidación requiere seguir avanzando en experimentación, calibración y transparencia, de modo que la personalización en mercados emergentes sea técnicamente viable, y, a su vez, social y comercialmente sostenible.

CONCLUSIONES

Este estudio demostró que un sistema de recomendación puramente histórico, implementado en dos etapas, generación de candidatos mediante co-ocurrencias/Item-CF y ordenamiento con BPR-MF, constituye una alternativa eficaz y viable para pymes de moda en contextos de alta rotación de catálogo y baja densidad de datos. Los resultados evidenciaron mejoras significativas en NDCG@k y Recall@k respecto a los modelos de popularidad, Item-CF y User-CF, además de un incremento en cobertura y diversidad intra-lista sin un deterioro relevante en la precisión.

La validación en usuarios mediante un experimento A/B corroboró que las mejoras offline se trasladan al entorno real, reflejándose en mayor CTR, menor tiempo de búsqueda y mejor percepción de usabilidad. Estos hallazgos confirman la validez del protocolo adoptado y la coherencia entre métricas técnicas y experiencia práctica. En términos prácticos, el enfoque histórico se muestra operativamente factible en infraestructuras de recursos moderados, favorece el descubrimiento de productos y aporta evidencia reproducible para la toma de decisiones en merchandising. Aunque se trata de un caso único que limita la generalización, el estudio sienta

las bases para futuras investigaciones orientadas a replicar y extender este protocolo, consolidando así a los sistemas basados en historial como una estrategia efectiva y transferible para la personalización en pymes de moda en mercados emergentes.

REFERENCIAS BIBLIOGRÁFICAS

1. Covington P, Adams J, Sargin E. Deep neural networks for YouTube recommendations. En: Proceedings of the 10th ACM Conference on Recommender Systems. Boston (MA): ACM; 2016. p. 191–8. https://dl.acm.org/doi/10.1145/2959100.2959190

2. Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L. BPR: Bayesian personalized ranking from implicit feedback. arXiv. 2012. http://arxiv.org/abs/1205.2618

3. He X, Liao L, Zhang H, Nie L, Hu X, Chua TS. Neural collaborative filtering. En: Proceedings of the 26th International Conference on World Wide Web. Perth (Australia): International World Wide Web Conferences Steering Committee; 2017. p. 173–82. https://dl.acm.org/doi/10.1145/3038912.3052569

4. Deldjoo Y, Nazary F, Ramisa A, McAuley J, Pellegrini G, Bellogin A, et al. A review of modern fashion recommender systems. ACM Computing Surveys. 2024;56(4):1–37. https://dl.acm.org/doi/10.1145/3624733

5. Ding Y, Lai Z, Mok PY, Chua TS. Computational technologies for fashion recommendation: a survey. ACM Computing Surveys. 2024;56(5):1–45. https://dl.acm.org/doi/10.1145/3627100

6. Duricic T, Kowald D, Lacic E, Lex E. Beyond-accuracy: a review on diversity, serendipity, and fairness in recommender systems based on graph neural networks. Frontiers in Big Data. 2023;6:1251072. https://www.frontiersin.org/articles/10.3389/fdata.2023.1251072/full

7. Zangerle E, Bauer C. Evaluating recommender systems: survey and framework. ACM Computing Surveys. 2023;55(8):1–38. https://dl.acm.org/doi/10.1145/3556536

8. del Aguila JR, Alva-Arévalo A, Cárdenas-García Á. Impact of e-commerce on business competitiveness and customer satisfaction. Diginomics. 2024;3:134.

9. Steck H. Calibrated recommendations. En: Proceedings of the 12th ACM Conference on Recommender Systems. Vancouver (Canadá): ACM; 2018. p. 154–62. https://dl.acm.org/doi/10.1145/3240323.3240372

10. Ríos del Aguila J, Alva-Arévalo A, Cárdenas-García Á. E-commerce and its relationship with customer satisfaction among Mishky Cacao association customers. Diginomics. 2024;3:117.

11. Abdollahpouri H, Nazari Z, Gain A, Gibson C, Dimakopoulou M, Anderton J, et al. Calibrated recommendations as a minimum-cost flow problem. En: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. Singapore: ACM; 2023. p. 571–9. https://dl.acm.org/doi/10.1145/3539597.3570402

12. Sato M. Calibrating the predictions for Top-N recommendations. En: 18th ACM Conference on Recommender Systems. Bari (Italia): ACM; 2024. p. 963–8. https://dl.acm.org/doi/10.1145/3640457.3688177

FINANCIACIÓN

Los autores no recibieron financiación para el desarrollo de la presente investigación.

CONFLICTO DE INTERESES

Los autores declaran que no existe conflicto de intereses.

CONTRIBUCIÓN DE AUTORÍA

Conceptualización: Jennifer Valeria Faz Gallardo, Valeria Anabel Guaman Capa.
Curación de datos: Rodolfo Najarro Quintero.
Análisis formal: Jennifer Valeria Faz Gallardo, Rodolfo Najarro Quintero.
Investigación: Rodolfo Najarro Quintero.
Metodología: Jennifer Valeria Faz Gallardo.
Administración del proyecto: Valeria Anabel Guaman Capa.
Recursos: Valeria Anabel Guaman Capa.

Software: Jennifer Valeria Faz Gallardo.
Supervisión: Rodolfo Najarro Quintero.
Validación: Rodolfo Najarro Quintero.
Visualización: Valeria Anabel Guaman Capa.
Redacción – borrador original: Jennifer Valeria Faz Gallardo, Valeria Anabel Guaman Capa, Rodolfo Najarro Quintero.
Redacción – revisión y edición: Jennifer Valeria Faz Gallardo, Valeria Anabel Guaman Capa, Rodolfo Najarro Quintero.

https://doi.org/10.62486/agma2025206 ISSN: 3046-4048